CSCI 699: The Science of Large Language Models

Units: 4.0

Fall 2024 MonWed 4:00-5:50PM

Location: TBD

Course website: https://robinjia.github.io/classes/fall2024-csci699.html . Please refer to the course website for up-to-date information.

Instructor: Robin Jia

Office: SAL 236

Office Hours: TBD

Contact Info: robinjia@usc.edu. I will reply within 48 hours. Please include “CSCI 699” in your email subject.

Course Description

Large language models (LLMs) are modern engineering marvels that have revolutionized natural language processing. Despite this success, there are still many open questions surrounding how and why LLMs work. This class will cover current research that considers LLMs as scientific objects of study. We will consider three complementary perspectives on understanding LLMs. First, we will analyze the internal operations of LLMs to shed light on how their predictions are computed. Second, we will study LLMs as black boxes and aim to discover principles that govern their behavior. Finally, we will survey external data-related factors that shape the general tendencies of LLMs. By understanding these different perspectives, students will develop a fuller understanding of modern research on LLMs.

Learning Objectives

Students will come away with a detailed understanding of large language models, including state-of-the-art literature that seeks to understand large language models from different scientific perspectives. Taking this class will prepare students for research on large language models. Students will get weekly practice analyzing and discussing current research papers.

Recommended Preparation

Familiarity with natural language processing (at the level of CSCI 544) or machine learning (at the level of CSCI 567). Email the instructor if you want to enroll but are unsure if you meet the recommended preparation.

Course Notes

Grading type: Letter or Credit/No Credit

Technological Proficiency and Hardware/Software Required

Students will be required to complete a final project that involves programming and running experiments with language models. No specific framework is required. Computing resources will be made available through CARC.

Required Readings and Supplementary Materials

All required readings will be provided in PDF form.

Optional Readings and Supplementary Materials

The course's recommended NLP textbook is Jurafsky and Martin, "Speech and Language Processing." The new 3rd edition is the most up-to-date NLP textbook available.

Description and Assessment of Assignments

Grades will be based on fulfilling roles in paper reading seminars (50%), general class participation (10%), and a final project (40% total).

Roles in Paper Reading Seminars (25% written + 25% in-class)

Students will take on various roles several times throughout the semester that will seed the in-class discussions on each paper. While the exact roles are subject to change, some potential roles include: Verbally describing the motivation and key research questions of the paper; Drawing a diagram that describes the main method(s) of the paper; Explaining the relationship between the paper and prior/subsequent work; Acting as "reviewer" of the paper as though it were a conference submission; Writing discussion questions for the paper; Writing reading comprehension quiz questions for the paper; and coming up with ideas for future work based on the paper. For each role, students must submit a 1 page report before class, and must also present their work during class to help initiate in-class discussions. Over the course of the class, each student will fulfill 10 different roles for different papers; each role will count for 5% of the overall grade, split evenly between the written report and in-class presentation.

This seminar format is heavily inspired by the format of Jesse Thomason's previous CSCI 699 class, which was in turn inspired by Alec Jacobson and Colin Raffel's blog post on Role-Playing Paper-Reading Seminars.

Class discussion participation (10%)

In addition to fulfilling their assigned roles, students are expected to participate in class discussions. This includes asking questions of the presenters and voicing opinions on discussion topics. Completion of in-class quizzes will also count towards the participation grade.

Final project (40% total)

Students must complete a final research project on a topic related to the class, either individually or in groups of two. This project is expected to include novel research that studies a scientific question about language models (which may or may not be "large," depending on resource constraints). While projects may involve querying closed-source models like ChatGPT, all projects must also study some open-source language models. Please come to office hours or email me if you have questions related to choosing a project direction.

Project proposal (5%). Students should submit a ~2-page proposal for their project by the end of Week 5. The proposal should describe the goal of the project and include a survey of related work.

Project midterm report (10%). Students should submit a ~4-page progress report for their project by the end of Week 10. This should once again describe the project's goals (which may have changed since the proposal), initial results, and a concrete plan of what will be done for the final report. While the initial results need not be positive, students are expected to have made non-trivial implementation progress by this point.

Project final presentation (10%). This will be a ~20 minute presentation during the last two weeks of class. Students should describe the motivation for their work, relevant background material, and results. I encourage students to present both positive and negative results. There will also be some time for audience questions.

Project final report (15%). Students should submit a ~6-page final report detailing all aspects of their project. The report should be structured like a conference paper, including an abstract, introduction, related work, and experiments. Parts of the proposal and progress report may be reused for the final report. Negative results will not be penalized, but should be accompanied with detailed analysis of why the expected results did not materialize.

Grading Breakdown

Table 1 Grading Breakdown

Assessment Tool (assignments)	**% of Grade**
Paper seminar written reports	25
Paper seminar presentations	25
Class participation	10
Project proposal	5
Project midterm report	10
Project final presentation	10
Project final report	15
TOTAL	100

Grading Scale

The course will use the following grading scale as a default:

Letter grade	Corresponding numerical point range
A	[93, ∞)
A-	[90, 93)
B+	[87, 90)
B	[83, 87)
B-	[80, 83)
C+	[77, 80)
C	[73, 77)
C-	[70, 73)
D+	[67, 70)
D	[63, 67)
D-	[60, 63)
F	[0, 60)

This grading scale may be altered only to lower these thresholds, i.e., only to make the final letter grades higher.

Assignment Submission Policy

Assignments should be submitted on Gradescope. All assignments will be due by 11:59pm on the due date. Written reports for paper seminar roles are due at 11:59pm on the day of the scheduled presentation. For additional information, see the course website.

Grading Timeline

Assignments will be graded within ten days of submission.

Course Specific Policies

Each student is given **5 late days** to use on any written assignment excluding the project final report (i.e., for any paper seminar role report, the project proposal, or the project midterm report). Additional late days will result in a deduction of 10% of the grade on the corresponding assignment per day. For students working in groups, submitting the project proposal or project midterm report 1 day late would require *all* students in the group to use 1 of their late days.

For the project proposal and midterm report, no late submissions will be accepted more than **3 days** after the stated due date.

Attendance

Students are expected to inform the course staff if they will be absent on a day when they are scheduled to fulfill a paper seminar role. Students who unexpectedly miss class may be allowed to make up a paper seminar role presentation by taking on an additional role later in the semester.

Academic Integrity

The University of Southern California is foremost a learning community committed to fostering successful scholars and researchers dedicated to the pursuit of knowledge and the transmission of ideas. Academic misconduct is in contrast to the university's mission to educate students through a broad array of first-rank academic, professional, and extracurricular programs and includes any act of dishonesty in the submission of academic work (either in draft or final form).

This course will follow the expectations for academic integrity as stated in the USC Student Handbook. All students are expected to submit assignments that are original work and prepared specifically for the course/section in this academic term. You may not submit work written by others or “recycle” work prepared for other courses without obtaining written permission from the instructor(s). Students suspected of engaging in academic misconduct will be reported to the Office of Academic Integrity.

Other violations of academic misconduct include, but are not limited to, cheating, plagiarism, fabrication (e.g., falsifying data), knowingly assisting others in acts of academic dishonesty, and any act that gains or is intended to gain an unfair academic advantage.

The impact of academic dishonesty is far-reaching and is considered a serious offense against the university and could result in outcomes such as failure on the assignment, failure in the course, suspension, or even expulsion from the university.

For more information about academic integrity see the student handbook or the Office of Academic Integrity’s website, and university policies on Research and Scholarship Misconduct.

Use of Generative AI in this Course

Since creating, analytical, and critical thinking skills are part of the learning outcomes of this course, all assignments should be prepared by the student working individually or in groups. Students may not have another person or entity complete any substantive portion of the assignment. Generative AI tools are trained, often without appropriate license, on text and images collected from the internet. Therefore, using AI generation tools is prohibited in this course unless explicitly marked as example outputs from such tools as part of an assessment or analysis of their behavior.

Course Content Distribution and Synchronous Session Recordings Policies

USC has policies that prohibit recording and distribution of any synchronous and asynchronous course content outside of the learning environment.

Recording a university class without the express permission of the instructor and announcement to the class, or unless conducted pursuant to an Office of Student Accessibility Services (OSAS) accommodation. Recording can inhibit free discussion in the future, and thus infringe on the academic freedom of other students as well as the instructor. (Living our Unifying Values: The USC Student Handbook, page 13).

Distribution or use of notes, recordings, exams, or other intellectual property, based on university classes or lectures without the express permission of the instructor for purposes other than individual or group study. This includes but is not limited to providing materials for distribution by services publishing course materials. This restriction on unauthorized use also applies to all information, which had been distributed to students or in any way had been displayed for use in relationship to the class, whether obtained in class, via email, on the internet, or via any other media. (Living our Unifying Values: The USC Student Handbook, page 13).

Course Schedule

Note: Reading materials are subject to change at the instructor’s discretion as the semester progresses. Refer to the course website for up-to-date information regarding the schedule and reading materials.

Table 3 Course schedule

	Topics/Daily Activities	Readings	Deliverables
Week 1	Introduction; Three Scientific Perspectives on LLMs	Vaswani et al., 2017	
Week 2	**Background**: LLM architectures. Transformers, the residual stream, tokenization, positional embeddings.	Sennrich et al., 2016; Su et al., 2021; Touvron et al., 2023	
Week 3	**Part 1**: Internals of LLMs. Localization, causal analysis of LLMs, model editing.	Dai et al., 2022; Meng et al., 2022; Geva et al., 2023; Hase et al., 2023;	
Week 4	Representations of meaning; representations of truth.	Li et al., 2021; Kim et al., 2023; Burns et al., 2022;	
Week 5	Circuits, induction heads, patching.	Elhage et al., 2021; Olsson et al., 2022; Wang et al., 2023; Merullo et al., 2024	Project Proposal due
Week 6	Logit lens, training dynamics, grokking.	Geva et al., 2022; Ghandeharioun et al., 2024; Nanda et al., 2023; Chen et al., 2023	
Week 7	**Part 2**: Analyzing Black-box LLM behavior. Memorization of training data.	Tirumala et al., 2022; Carlini et al., 2023; Chang et al., 2023	
Week 8	In-context learning: When does it work? Why does it work?	Min et al., 2022; Yoo et al., 2022; Wang et al., 2022.	
Week 9	Chain-of-thought reasoning, faithfulness, verbalizing confidence.	Turpin et al., 2023; Zhou et al., 2023.	
Week 10	**Part 3**: External forces and the role of data. Influence of pre-training data, data deduplication, filtering.	Lee et al., 2022; Elazar et al., 2023; Zhang et al., 2023; Park et al., 2023.	Project Midterm Report due
Week 11	Scaling laws, emergent abilities. Are emergent abilities a mirage?	Kaplan et al., 2020; Hoffmann et al., 2022; Wei et al., 2022; Schaeffer et al., 2023	
Week 12	Measuring LLM "opinions," effects of fine-tuning.	Santurkar et al., 2023; Zhou et al., 2023; Jain et al., 2023; Ryan et al., 2024.	
Week 13	Feedback loops in data; LLMs as LLM evaluators.	Taori et al., 2022; Seshadri et al., 2023; Shen et al., 2023; Boyeau et al., 2024.	
Week 14	Final project presentations		
Week 15	Final project presentations		
FINAL	Final Project Report		Project Final Report due on the university-scheduled date of the final exam.

Statement on Academic Conduct and Support Systems

Academic Integrity:
The University of Southern California is a learning community committed to developing successful scholars and researchers dedicated to the pursuit of knowledge and the dissemination of ideas. Academic misconduct, which includes any act of dishonesty in the production or submission of academic work, compromises the integrity of the person who commits the act and can impugn the perceived integrity of the entire university community. It stands in opposition to the university's mission to research, educate, and contribute productively to our community and the world.

All students are expected to submit assignments that represent their own original work, and that have been prepared specifically for the course or section for which they have been submitted. You may not submit work written by others or "recycle" work prepared for other courses without obtaining written permission from the instructor(s).

Other violations of academic integrity include, but are not limited to, cheating, plagiarism, fabrication (e.g., falsifying data), collusion, knowingly assisting others in acts of academic dishonesty, and any act that gains or is intended to gain an unfair academic advantage.

The impact of academic dishonesty is far-reaching and is considered a serious offense against the university. All incidences of academic misconduct will be reported to the Office of Academic Integrity and could result in outcomes such as failure on the assignment, failure in the course, suspension, or even expulsion from the university.

For more information about academic integrity see the student handbook or the Office of Academic Integrity's website, and university policies on Research and Scholarship Misconduct.

Please ask your instructor if you are unsure what constitutes unauthorized assistance on an exam or assignment, or what information requires citation and/or attribution.

Students and Disability Accommodations:
USC welcomes students with disabilities into all of the University's educational programs. The Office of Student Accessibility Services (OSAS) is responsible for the determination of appropriate accommodations for students who encounter disability-related barriers. Once a student has completed the OSAS process (registration, initial appointment, and submitted documentation) and accommodations are determined to be reasonable and appropriate, a Letter of Accommodation (LOA) will be available to generate for each course. The LOA must be given to each course instructor by the student and followed up with a discussion. This should be done as early in the semester as possible as accommodations are not retroactive. More information can be found at osas.usc.edu. You may contact OSAS at (213) 740-0776 or via email at osasfrontdesk@usc.edu.

Support Systems:

Counseling and Mental Health - (213) 740-9355 – 24/7 on call
Free and confidential mental health treatment for students, including short-term psychotherapy, group counseling, stress fitness workshops, and crisis intervention.

988 Suicide and Crisis Lifeline - 988 for both calls and text messages – 24/7 on call
The 988 Suicide and Crisis Lifeline (formerly known as the National Suicide Prevention Lifeline) provides free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week, across the United States. The Lifeline is comprised of a national network of over 200 local crisis centers, combining custom local care and resources with national standards and best practices. The new, shorter phone number makes it easier for people to remember and access mental health crisis services

(though the previous 1 (800) 273-8255 number will continue to function indefinitely) and represents a continued commitment to those in crisis.

Relationship and Sexual Violence Prevention Services (RSVP) - (213) 740-9355(WELL) – 24/7 on call
Free and confidential therapy services, workshops, and training for situations related to gender- and power-based harm (including sexual assault, intimate partner violence, and stalking).

Office for Equity, Equal Opportunity, and Title IX (EEO-TIX) - (213) 740-5086
Information about how to get help or help someone affected by harassment or discrimination, rights of protected classes, reporting options, and additional resources for students, faculty, staff, visitors, and applicants.

Reporting Incidents of Bias or Harassment - (213) 740-5086 or (213) 821-8298
Avenue to report incidents of bias, hate crimes, and microaggressions to the Office for Equity, Equal Opportunity, and Title for appropriate investigation, supportive measures, and response.

The Office of Student Accessibility Services (OSAS) - (213) 740-0776
OSAS ensures equal access for students with disabilities through providing academic accommodations and auxiliary aids in accordance with federal laws and university policy.

USC Campus Support and Intervention - (213) 740-0411
Assists students and families in resolving complex personal, financial, and academic issues adversely affecting their success as a student.

Diversity, Equity and Inclusion - (213) 740-2101
Information on events, programs and training, the Provost's Diversity and Inclusion Council, Diversity Liaisons for each academic school, chronology, participation, and various resources for students.

USC Emergency - UPC: (213) 740-4321, HSC: (323) 442-1000 – 24/7 on call
Emergency assistance and avenue to report a crime. Latest updates regarding safety, including ways in which instruction will be continued if an officially declared emergency makes travel to campus infeasible.

USC Department of Public Safety - UPC: (213) 740-6000, HSC: (323) 442-1200 – 24/7 on call
Non-emergency assistance or information.

Office of the Ombuds - (213) 821-9556 (UPC) / (323-442-0382 (HSC)
A safe and confidential place to share your USC-related issues with a University Ombuds who will work with you to explore options or paths to manage your concern.

Occupational Therapy Faculty Practice - (323) 442-2850 or otfp@med.usc.edu
Confidential Lifestyle Redesign services for USC students to support health promoting habits and routines that enhance quality of life and academic performance.